**Project-Exam 1 Deadline on 6 July 2020 11:59 PM**

1. It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

The link [here](https://umkc.box.com/s/6cnan5zesgntmsxzlgbjwjoiii6fywmj) is a dataset for fraud detection Link.

**Description of the dataset**

**1)Time**: Number of seconds elapsed between this transaction and the first transaction in the dataset

**2)V1-V28**: Result of a PCA Dimensionality reduction to protect user identities and sensitive features

1. Apply Any classification of your choice (KNN, Naïve Bayes, SVM, Random Forest, …) and report the performance
2. Visualize the number of samples per class (This is a binary classification, 0: Non-Fraud and 1: Fraud) and report your observation
3. This dataset is unbalanced meaning we don’t have equal number of samples per class. Consequently, we need specific techniques when dealing with **unbalanced dataset**. Please study one of the techniques or challenges we face while working on unbalanced dataset and discuss it shortly.

1. Apply K-means on the dataset in this [link](https://umkc.box.com/s/lwr3s70prbe3tdifzx6fblq7wqln54xq) and visualize the clusters using matplotlib or seaborn.

**Description of the dataset:**

**1)Customer\_id**

**2)Age**

**3)Annual Income**

**4)Spending score**

## **Our goal is to cluster our customers into buying groups based off of their Annual Income and Spending Scores**

a. Report which K is the best using the elbow method.

b. Evaluate with silhouette score or other scores relevant for unsupervised approaches (before applying clustering clean the data set if needed)

c. Can you interpret the clustering result that you have visualized?

1. Use the dataset in [this](https://umkc.box.com/s/de9bdscv3ys9ff1sjogbfzppzy1o477a) link. Predict the temperature using the weather details specified in the columns

**Description of the dataset:**

This dataset has 12 columns and we want to predict “Temperature (C)” using other independent variables.

a) Apply some Exploratory Data Analysis to draw some insight from the data

a) Visualize the data and draw the model line

b) Evaluate the model and try to interpret the performance that you get

1. Use the dataset in [this](https://umkc.box.com/s/mlmyznn0667tbgui5urnzbos1i1c48qf) link and apply classification on that.

**Description of the dataset:**

The dataset is Spam Dataset and has two columns

**1) class**

**2) text**

You need to initially clean the text data and apply the techniques we have learned for transforming the text data into numeric format (TFIDF, Count\_Vectorizer, …)

Evaluate the model and try to interpret the result

5) Pick any dataset online for the classification problem which includes both numeric and non- numeric features

* 1. Perform exploratory data analysis on the data set (it can be anything on your choice that gives insight about the dataset)
  2. Apply the three classification algorithms Naïve Bayes, SVM and KNN on the chosen data set and report which classifier gives better result.
  3. Try SVM with linear and non-linear kernel and report which one gives better performance

**Submission Guidelines:**

1. submission must be in a group of three or four students.

2. Submit your source code and documentation to GitHub and represent the work through wiki page properly (submit your screenshots as well. The screenshot should have both the code and the output)

3. Comment your code appropriately

4. Video Submission (2 –3 min video showing the demo of the project-exam, with a brief voice over on the code explanation)

5. Submit only report at Turnitin in UMKC

6. Remember that similarity score should be less than 15%

7. **Submit your work through the feedback form to the github**

7. The report should include below details

I. Introduction

II. Objectives

III. Approaches/Methods

IV. Workflow

V. Datasets (if applicable)

VI. Parameters

VII. Evaluation & Discussion

VIII. Conclusion

**Evaluation Criteria:**

1. Report similarly score (should be less than 15%)

2. Report Quality (check the below example reports for reference)

3. Time (should submit before due time)

4. Wiki page Example Reports:

<https://github.com/stratospark/food-101-keras>

<https://github.com/matterport/Mask_RCNN>

<http://blog.stratospark.com/deep-learning-applied-food-classification-deep-learning-keras.html>

**University Policy on Student Conduct:**

Cheating, plagiarism, disruptive behavior and other forms of unacceptable conduct are subject to strong sanctions in accordance with university policy. See detailed description of university policy at the following URL:

<https://catalog.umkc.edu/special-notices/academic-honesty/>

<https://www.umsystem.edu/ums/rules/collected_rules/programs/ch200/200.010_standard_of_conduct>